Dictionnaire intelligent d'aide à la compréhension

Xabier Artola Zubillaga – Fabrice J. Evrard

1. Vue générale du projet

1.1. Introduction

Dans la recherche du traitement du langage natural, l'aspect du lexique devient de plus en plus prépondérant. En effet, il s'agit d'un champ de recherche dont on a fait bon marché autrefois mais qui s'est révélé d'une grande importance à l'heure de construire de vrais systèmes de comprehénsion du langage naturel doués de composantes lexicales de taille réelle.

La construcción à la main de ces composantes lexicales de taille réelle est bien sûr un travail énorme. Par ailleurs, il semble évident que dans les éditions courantes des dictionnaires on peut trouver de grandes quantités d'information qui —convenablement traitées— peuvent s'avérer être un point de départ important pour la construction automatique ou semi-automatique de ces bases lexicales. Ce traitement sera fait sur les versions machine (Machine Readable) de ces dictionnaires mais il ne faut pas oublier que ces versions, dans la plupart des cas, ne sont que des bandes magnétiques expressément enregistrées dans le but de servir de source aux versions papier correspondantes. C'est pour cela que le traitement automatique des MRD, afin d'y extraire toutes les informations possibles sur les mots, n'est pas non plus un travail trivial. Le livre de Boguraev et Briscoe (1989) présente le champ de recherche de la lexicographie computationnelle et montre plusieurs travaux realisés ou en cours.

Notre travail diverge fondamentalement des travaux mentionnés ci-dessus par le type des utilisateurs pour qui on a conçu le système: il s'agit d'un système dictionnarial monolingue pour des usagers humains.

1.2. Dictionnaire intelligent d'aide à la compréhension

Aujourd'hui on dispose déjà de dictionnaires electroniques grâce aux grandes possibilités que nous offre la technologie des CD-ROMs qui se révele d'une puissance énorme pour l'emmagasinage de grandes quantités d'information. Le problème qui se pose immédiatement avec ces outils est évidemment celui de l'accès à toute cette information: on peut bien concevoir un dictionnaire electronique dans lequel toutes les voies d'accès aux données sur les mots soient limitées à l'accès alphabétique traditionnel, c'est-à-dire que l'utilisateur n'aura pas d'autres voies alternatives de recherche de l'information. Cela est vraiment pauvre. C'est pour cela qu'il faut construire des interfaces vraiment puissantes, des interfaces capables d'exploiter toute cette connaissance de façons diverses et enfin des interfaces qui vont se comporter de manière intelligente lors de l'accès au dictionnaire.

Le système que nous sommes en train de construire est une interface intelligente à un dictionnaire monolingue du français qui va servir d'aide à la compréhension

—en général par les diverses fonctions qu'un tel dictionnaire remplit dans les mains d'un usager.

Il s'agit d'un problème double:

— Problème de la représentation de la connaissance contenue dans un dictionnaire monolingue: il s'agit de représenter justement les concepts contenus dans le dictionnaire, les attributs associés à ces concepts, les rapports entre ces concepts.
— Exploitation intelligente de toute cette information: il faut que le système soit capable de fonctionner de manière analogue à un utilisateur qui se sert d'un dictionnaire intelligemment quand il extrait des informations non explicitement contenues mais inférables des données explicites. Pour cela il faut donner au système des capacités d'accès diverses et des mécanismes d'inférence puissants.

C'est au travers de la conception et l'élaboration d'un dictionnaire informatique que l'on devra dégager une architecture logicielle devant aboutir à l'implantation d'un exemplaire, éventuellement remanié, d'un dictionnaire monolingue. Il devra être tel que tout usager pourra, de façon conviviale mais surtout intelligente, dialoguer avec ce système afin de mieux appréhender la signification des mots, les modes d'expression du sens et plus généralement de mieux posséder la langue dans tout ce que cela implique de savoir faire.

Parmi les buts du travail de recherche entrepris on a celui d'étudier les méthodes d'élaboration des dictionnaires monolingues utilisées par les lexicographes afin de comprendre les liens jugés pertinents entre concepts et de proposer une représentation adéquate, permettant les raisonnements habituels et naturels du lecteur. Il s'agit bien ici de définir les comportements souhaités d'un lecteur-utilisateur d'un tel dictionnaire et de les confronter au mode de représentation et d'utilisation choisis.

Dans ce but, on ne parlera en avant que de la composante sémantique du dictionnaire monolingue, c'est-à-dire des définitions.

La première phase du projet consiste donc à construire ce que nous appelons la **base de connaissances dictionnariale**. Une grande partie des informations dans cette base de connaissances va être constituée à partir de la version imprimée d'un dictionnaire, ou mieux, de la version machine de celui-ci; c'est un travail d'interprétation et de reformulation des données qu'il est inconcevable de réaliser manuellement. Il faut donc essayer d'automatiser tout ce qui est susceptible de l'être. Cependant, comme de nombreux auteurs le signalent, on ne pourra pas tout automatiser et il faudra fournir des procédures sémi-automatiques et même manuelles pour réaliser quelques tâches concernant cette reformulation.

Nous pensons ainsi que ce travail d'interprétation et de reformulation des textes d'un dictionnaire, dont le but est d'obtenir une représentation disons exploitable des connaissances qui y sont contenues, va être réalisé comme suit:

— *automatiquement:* il faudra avoir un analyseur syntaxique/sémantique qui, étant donnée une définition, construise la structure sémantique correspondante; cet analyseur devra être robuste de telle sorte qu'il n'échoue pas complétement quand il trouve quelque chose d'inattendu et qu'il construise partiellement la structure correspondante

— *sémi-automatiquement:* il s'agit du cas où l'analyseur mentionné n'est pas capable de construire para lui-même la structure sémantique correspondant à la phrase de définition et il demande l'aide de l'opérateur; il va être dans le cas du lexicographe qui, contraint par les possibilités offertes par le système, va reformuler d'une manière adéquate la définition.

Le procéssus de construction de la base de connaissances dictionnariale va se débrouiller de manière interactive, ayant, dans cette première phase de construction, au moins les fonctions suivantes:

— Analyse syntaxique-sémantique des phrases de définition du MRD avec les caractéristiques mentionnées ci-dessus, pour obtenir la représentation sémantique désirée dans la BCD.
— Interaction avec le lexicographe afin de résoudre le cas non résolvables sans intervention humaine, en lui présentant les possibilités qu'il peut utiliser.
— Déduction des faits déductibles au fur et à mesure qu'elle se construit: désambiguïsation de sens, etc.
— Vérification de la consistance de la BCD.

1.3. Plan de l'article

Dans ce qui suit nous allons montrer les études réalisées sur un petit dictionnaire monolingue du français en ce qui concerne l'etiquetage des occurrences des mots dans les textes de définition (fondamentalement avec leur catégorie grammaticale) et leur désambiguïsation partielle. Tout cela a été utilisé pour obtenir une série de schémas de définition pour les catégories les plus nombreuses (substantifs, adjectifs et verbes) que nous utilisons après pour proposer un formalisme de représentation sémantique et comme justification de celui-ci. De même on présente ici les relations sémantiques dégagées de ces schémas de définition comme la synonymie, hyponymie et hyperonymie, relations meronymiques, etc. et les divers mots-fonction que le lexicographe utilise quand il définit d'autres mots. Pour finir nous proposons un formalisme de représentation et faisons un ébauche de l'aspect déductif du système implanté pour le moment.

2. Travail empirique réalisé sur un dictionnaire réel

Il s'agit d'étudier quelle est l'information contenue dans un dictionnaire monolingue explicatif. Cela dépend bien sûr de beaucoup de facteurs: qualité du dictionnaire, taille, public visé, etc. Outre cela, nous pouvons trouver dans un dictionnaire de ce type des informations très variées: phonétique, orthographique, morphosyntaxique, éthymologique, sur l'usage, sémantique, pragmatique, etc. Mais ce qui nous intéresse surtout, comme on l'a déjà dit, c'est l'aspect sémantique du dictionnaire, le dictionnaire comme *conteneur d'un ensemble de sens ou acceptions en interrelation.* Dans un dictionnaire du type mentionné ci-dessus on va chercher cela dans les définitions des mots.

Donc il faut étudier ces définitions. Pour cela, nous nous sommes fixés sur *Le*

Plus Petit Larousse, LPPL (1980, Librairie Larousse, Paris), un petit dictionnaire monolingue du français qui nous a permis de faire quelques études statistiques sur la fréquence et la nature grammaticale des mots. Ces études nous montrent, par exemple, certains mots qui appartiennent au métalangage définitoire des mots, c'est-à-dire des mots que le lexicographe utilise fréquemment lorsqu'il définit un concept.

Comme suite à ces études statistiques on est arrivé à faire un recensement des schémas de définition dont une typologie doit en principe nous conduire à adopter une représentation de ces connaissances. Un tel choix impliquera nécessairement des modes de manipulation qui rendront ce dictionnaire intelligent.

2.1. Caractéristiques du LPPL

Pour continuer nous allons décrire ces études et montrer quelques données que l'on peut en tirer. Tout d'abord il faut dire que, faute d'une version machine du LPPL on a rentré ce dictionnaire sur un système de gestion de base de données relationnelle dont on se servira désormais pour les études qu'on va décrire. Le LPPL posséde 15953 entrées (dont 70.09% sont des entrées monosémiques et 21.29% des entrées à deux acceptions) avec un nombre de 22899 sens et 1980 formes dérivées (féminins des adjectifs, surtout). Pour chaque entrée il nous donne information sur l'orthographe du mot, catégorie grammaticale (en la marquant aussi pour chaque sens qui a une catégorie différente), et dans un petit nombre de cas marque d'usage (*Mus.*, *Fam.*, etc.), exemples (14% des cas) et d'autres informations additionnelles: conjugaison des verbes, phonétique, etc. Les définitions sont extremement courtes en général (3.27 mots/définition, moyenne), dont 74.57% ont entre 1 et 4 mots. On peut trouver 13740 sens qui appartiennent à la catégorie des noms, 5259 verbes, 3731 adjectifs et 173 adverbes.

2.2 Étude fréquentielle: classement preliminaire des mots

La première étude statistique qu'on a effectué sur LPPL a été une étude fréquentielle sur les mots appartennant aux définitions. Ces études statistiques montrent que ce sont les mots dits grammaticaux (prépositions, conjonctions, etc.) qui ont les plus hautes fréquences. Parmi ces mots on peut constater qu'il y a des différences notables selon la catégorie à laquelle ils appartiennent: par exemple, *de* c'est la forme canonique la plus fréquente (10.57%) parmi les mots des définitions des noms tandis que parmi les définitions des verbes c'est *qui* la plus utilisée (8.80%). Cela suggère sans doute l'existence de schémas du style ‹NOM› *de* ‹COMPLEMENT›, c'est-à-dire des définitions classiques du type «genus et differentia» pour les noms. Pour les adjectifs c'est le schéma *qui* ‹GROUPE VERBAL› qui est suggeré, le groupe verbal indiquant l'action réalisée ou subie par ce qui est attribué par l'adjectif défini.

A part ces relativement hautes fréquences des mots grammaticaux, on peut constater aussi des différences notables entre les catégories grammaticales des mots de définition parmi les définitions appartenant aux différents sous-ensembles des entrées: par exemple, dans les définitions des noms on trouve des substantifs parmi les mots avec les plus hautes fréquences: mots comme *action, partie, personne, plante, chose, corps, sorte, ensemble, état, instrument,* etc., c'est-à-dire les concepts qui vont se pla-

cer sur les niveaux les plus élevés dans les hiérarchies des concepts, ceux qui ont beaucoup d'hyponymes; dans les définitions des verbes par contre on ne trouve pas des noms avec de fréquences importantes mais des verbes à l'infinitif, voire *faire, rendre, mettre, être, donner, avoir, prendre* (verbes support, relateurs, souvent) en indiquant les genera les plus importants pour les verbes. L'étude sur les définitions des adverbes, pronoms, articles, conjonctions, etc. dans le dictionnaire montre une autre manière pour définir ce genre d'entrées: le lexicographe utilise le métalangage grammatical dans ces cas comme le montre le fait qu'on trouve parmi les mots de plus haute fréquence des mots comme *temps, quantité, lieu, façon, négation, adv.*, etc. Pour finir, la présence relative de nombreuses formes conjugées des verbes comme *a, peut, est, dit, fait* dans les définitions des adjectivs suggère aussi la structure de groupe verbal dont on vient de parler.

Ces études statistiques nous ont amené à une première classification —bien que provisoire— des mots intervenant dans les définitions de ce petit dictionnaire en vue de son utilisation postericure dans le formalisme de représentation des connaissances qui sera retenu.

2.3. Etiquettage (tagging) et désambiguïsation

Le travail de désambiguïsation des occurrences des mots dans les définitions du dictionnaire a consisté à leur ajouter un numéro de sens qui va distinguer chacune des occurrences des autres acceptions du même mot. Cela vient de la nécessité d'arriver à un réseau où les noeuds ne représenteront pas des mots mais des sens différents de ces mots. Par ailleurs, l'étiquettage ou «tagging» consiste essentiellement en l'affectation de chaque occurrence des mots dans les textes de définition avec sa catégorie grammaticale correspondnate. Naturellement, aussi bien pour l'etiquettage que pour la désambiguïsation on s'est servi dans un premier temps des informations contenues dans le LPPL lui-même.

Un premier problème s'est posé au moment de travailler sur des occurrences qui ne sont pas des entrées du dictionnaire (21373 occurrences exactement, presque 30% du total); après avoir examiné cet ensemble de mots on a vu qu'il s'agit principalement de mots assez fréquents dans les définitions qui manquent dans le LPPL, des altérations morphologiques régulières des noms, adjectifs et verbes, des altérations orthographiques —notamment apostrophés et abrégés—, des signes de ponctuation, des erreurs de transcription et d'autres. En vue de réduire ce nombre d'occurrences de mots qui n'étaient pas susceptibles d'étiquettage ou de désambiguïsation automatiques on a procédé de la façon suivante:

1. On a décidé d'ajouter certaines entrées (31 mots, 73 en nombre de sens) à la base de données en tirant leurs définitions d'un autre petit dictionnaire (*Larouse de Poche*, Librairie Larousse, 1979. Paris) pour les cas les plus flagrants des entrées qui manquent dans le LPPL. On a parmi eux des mots comme *au, course, et, non, pas* (adv.), *voyager*, etc.
2. On a ajouté à chaque occurrence du LPPL la forme canonique correspondante. Ce travail a été fait à la main dans les cas douteux des occurrences les plus fréquentes (fréquence d'apparition > 10) en résolvant ainsi les possibilités d'analyses multiples qu'un analyseur morphologique nous donne-

rait: l'occurrence *fait*, par exemple, a été affectée avec des formes canoniques *fait* (substantif) ou *faire* selon la fonction qu'elle remplissait dans chaque cas. Pour le reste des cas et pour l'instant, on a fait de telle sorte que la forme canonique soit égale à l'occurrence.

3. On a tiré profit aussi des mots dérivés qui apparaîssent comme entrées du LPPL et qui avaient été enregistrés: il s'agit surtout des féminins de quelques adjectifs.

Après cela nous sommes arrivés à réduire le nombre de formes canoniques sans possibilités d'être étiquettées automatiquement à 4770 avec un nombre d'occurrences de 8172 (un peu superieur à 10% du total d'occurrences).

Une fois passé le programme de étiquettage-désambiguïsation on a constaté que certains mots parmi les plus fréquents restaient encore non désambiguïsés, et même non étiquettés avec leur catégorie, car figurant come entrées polysémiques de plusieurs catégories: par exemple *en* «prép.» et «pron. pers.». C'était très ennuyeux pour obtenir des schémas syntaxiques des définitions car beaucoup de celles-ci restaient ambiguës à cause d'un de ces mots-ci; il a alors fallu désambiguïser certains d'entre eux à la main et d'autres en utilisant des heuristiques fondées sur des critères de type de langage utilisé dans les définitions, de co-occurrence de différentes catégories, etc.

Lorsqu'on a obtenue une première liste de schémas syntaxiques des définitions on en a utilisée pour encore étiqueter des mots par analogie des schémas: par exemple, dans les définitions avec schéma ‹?› ‹PREP.› ‹NOM› on a etiqueté le premier mot comme NOM car la plupart des définitions homologables (pour le cas des substantifs) étaient ‹NOM› ‹PREP.› ‹NOM› (701 vs. 7 ayant des schémas différents).

Les pourcentages d'étiquettage (désambiguïsation par rapport à la catégorie) et de désambiguïsation «complète» (affectation du sens correspondant en plus de la catégorie) obtenus après le procéssus réalisé sont les suivants:

— 64716 occurrences de mot (84.8%) étiquettées avec leur catégorie grammaticale (dont 26032, 34.11%, possèdent déjà la spécification du sens correspondant: il s'agit évidemment dans la plupart des cas d'occurrences de mots monosémiques dans le LPPL)
— 15508 définitions, 67.45% des sens, complétement étiquetées (catégorie ajoutée dans toutes les occurrences appartennant)

Il faudrait maintenant utiliser un petit analyseur morphologique afin de résoudre la plupart des cas restants en rapport avec l'étiquettage de catégorie (en fait il s'agit notamment des cas au pluriel); la désambiguïsation du sens présente des problèmes plus difficiles qui vont exiger des procédures sémi-automatiques pour la resoudre complètement.

2.4. Schémas de définition syntaxiques

Les schémas de définition obtenus sont des schémas conçus comme des «séquences de catégories» correspondant aux définitions totalement étiquettées dans le processus de «tagging». On ne parle pas ici des schémas éventuellement dégageables en fonction des mots spécifiques du métalangage définitoire mais des schémas purement syn-

taxiques auxquels on va essaier ensuite d'associer des représentations sémantiques adéquates.

Dans ce qui suit nous présentons comme exemple une petite grammaire en forme EBNF des définitions des noms: il s'agit évidemment d'une structure de groupe nominal et selon nos estimations entre 60 et 85% des définitions des noms dans le LPPL la suivent.

définition_de_nom = *[qualificatif]* ***nom*** *[modificateur] [","* *définition_de_nom] | proposition_relative.*
modificateur = *qualificatif | préposition [quantificateur] [qualificatif] nom [modificateur] | préposition infinitif_verbal | [préposition pron_dém] proposition_relative.*

3. Représentation sémantique

3.1. Synonymie

La définition d'un mot qu'utilise un ou plusieurs synonymes (ou quasi-synonymes) constitue une technique utilisée souvent dans les dictionnaires. Évidemment, cette technique est encore plus utilisée dans les petits dictionnaires.

Ces liens synonymiques sont la plupart du temps implicites dans les définitions où un seul mot de la même catégorie du défini est utilisé pour le définir. On a aussi des définitions avec référence explicite de la relation synonymique *("syn. de...")* mais cela est très rare dans le LPPL.

Il s'agit principalément des schémas du style *nom [","* *nom]* pour les définitions des substantifs, et pareil pour les verbes et adjectifs. Parfois, ce genre de définition peut être adjointe d'un autre type de définition; par exemple, *nom [","* *définition_de_nom]* pour un substantif, où la deuxième partie de la définition n'est pas de type synonymique.

3.2. Relations taxonomiques: hyperonymie et hyponymie

Les liens taxonomiques de type hyperonymique (et donc son inverse hyponymique) seront établis surtout à partir des définitions classiques du type «genus et differentia». Ce sont les définitions où le noyau («genus») est un mot de la même catégorie grammaticale que le mot défini et il est accompagné d'un modificateur ou complément qui constitue la partie spécifique ou differentielle. Le «genus» représente l'hyperonyme du concept défini.

Sur la première ligne de la petite grammaire décrite dans la section 2.4 on peut trouver le noyau, qui est un nom pour les cas des définitions des noms: il est éventuellement accompagné de modificateurs ou compléments qui constituent le «differentia».

Comme Amsler (1980, 81, 84) et beaucoup d'auteurs l'indiquent on va avoir des «tangled hierarchies» de concepts (par forcément des structures arborescentes) qui vont nous servir pour exploiter des propriétés héritées des niveaux supérieurs aux inférieurs.

Cependant, toutes les définitions répondant à ces schémas ne vont pas établir des liens taxonomiques de ce genre entre le mot défini et le noyau syntaxique de sa définition. D'après certaines études (Vossen et alt., 1989) on a vu que fréquemment le noyau syntaxique d'un texte de définition ne le correspond pas avec le concept le plus important du point de vue sémantique. Parfois, on peut retrouver des cas où en plus d'un lien taxonomique avec le noyau syntaxique on peut aussi, grâce aux caractéristiques du métalangage définitoire utilisé par le lexicographe, dégager une autre relation sémantique —plus spécifique et sûrement plus intéressante— avec un autre élément de la définition. C'est ce qu'on voudrait faire remarquer dans la section 3.4 en présentant quelques mots et structures qu'on a appelé mots-fonction.

Par ailleurs, il faut considérer aussi dans les relations taxonomiques la propre relation de taxonomie établie grâce à mots-fonction spécifiques et suivant cette syntaxe:

("espèce de" | "sorte de" | genre de") groupe_nominal

3.3. Relations meronymiques

Les relations meronymiques (en général, part-whole) sont dégageables aussi des définitions des noms du dictionnaire. Elles répondent aux schémas suivants:

[qualificatif] "partie" [qualificatif] "de" groupe_nominal.
("élément de" | "membre de") groupe_nominal
"pièce de" groupe_nominal
("ensemble de" | "reunion de" | "groupe de") groupe_nominal_pluriel
nom ("à" | "avec") nom (dans certains cas)

Pour l'exploitation intelligente des propriétés meronymiques il faudra tenir compte des problèmes que pose leur transitivité (Winston et alt., 1987).

3.4. Mots fonction dégagés

Tout d'abord, c'est bien clair à notre avis que la présence relativement importante de certaines relations sémantiques comme la synonymie, par exemple, est due surtout à des caractéristiques spécifiques du dictionnaire étudié: le LPPL est un très petit dictionnaire de poche et le lexicographe s'est servi souvent de la synonymie pour définir des concepts. De même, les mots fonction qu'on peut dégager de l'étude d'un seul dictionnaire répondent surtout à des critères particuliers du lexicographe au moment de rédiger les définitions et à une certaine tradition lexicographique; bien que cela ne donne à ces mots aucune universalité, nous pensons que les rapports qu'ils établissent entre les concepts peuvent aisément se trouver dans d'autres dictionnaires —du français même ou d'autres langues— avec des réalisations superficielles différentes.

Le choix de ces mots a été effectué par des critères statistiques: on s'est fixé sur des mots avec des fréquences relativement hautes dans les textes de définition et on a procédé à l'étude de leurs occurrences. La liste qu'on va présenter ne cherche pas

à être exhaustive. Ensuite nous présentons le classement qu'on a fait de ces mots en donnant la syntaxe de leurs réalisations dans le LPPL.

Déviateurs ou «shunters» (Vossen et alt., 1989): il s'agit de constructions du métalangage définitoire qui mettent en rapport le mot défini avec un mot d'une autre catégorie grammaticale —en général— qui porte le poids sémantique de la définition.

ACTION-DE	*[déterminant] "action de" infinitif_verbal [complément].*
ÉTAT-DE	*"état de ce qui est" (qualificatif | participe_verbal).*
QUALITÉ-DE	*"qualité de" /"ce qui est"/ qualificatif.*
RELATIF-A	*(("relatif à" | "de" | "propre" ("à" | "de")) [déterminant]) | "du" | "des" nom.*
FACULTÉ-DE	*"faculté de" infinitif_verbal.*
CARACTÈRE-DE	*"caractère de ce qui est" qualificatif.*
MANIÈRE-DE	*"manière de" infinitif_verbal.*
CONFORME-À	*"conforme à" [déterminant] nom.*
CONTRAIRE-À	*"contraire à" [déterminant] nom.*
RENDRE	*"rendre" [graduateur] (qualificatif | participe_verbal).*
CE-QUI	*"ce qui" groupe_verbal.*
CE-QUE	*"ce que" groupe_verbal.*
QUI	*"qui" groupe_verbal.*
SANS	*"sans" nom.* (défs. des adjectifs)

Autres relateurs:

SON	*"son" nom.*
SE	*"se" infinitif_verbal [complément].*
NON	*"non" qualificatif.*
MANQUE-DE	*"manque de" nom.* (défs. des noms)

3.5. Différentes structures de configurations. Sémantique attachée

Pour la représentation du signifié des textes de définition on va se fonder sur les diverses structures syntaxiques décrites avant. Les structures à base de mots-fonction vont nous fournir des constructions sémantiques spécifiques du langage dictionnarial; les études préalables sur les «déviateurs» et autres relateurs nous ont permis d'établir un ensemble de liens entre concepts —défini et définissants, ou même entre constituants de la définition— dont il faut profiter à notre avis pour la représentation sémantique des textes de définition d'un dictionnaire.

D'un autre côté, l'étude sur la syntaxe de ces textes de définition nous montre que, pour un grand nombre de cas, il est possible de décrire leur syntaxe au moyen d'un nombre relativement petit de règles. Mis à part les cas des relateurs spécifiques mentionnés, on va se fonder sur ces structures syntaxiques pour l'établissement de liens entre concepts.

La structure de groupe nominal dans les définitions des substantifs va nous permettre de trouver le nom hyperonyme du nom défini comme noyau syntaxique du texte de définition et des attributs attachés à celui-ci faisant partie des divers modifi-

cateurs présents. De même, on va trouver l'hyperonyme d'un verbe comme noyau syntaxique —verbe à l'infinitif— du groupe verbal constituant sa définition et on va tirer profit d'une représentation du style grammaire casuelle pour représenter les rapports entre le noyau et les autres constituants de la définition. Le cas des adjectifs qualificatifs va être différent car la plupart de leurs définitions ne nous amènent pas aux hyperonymes mais elles expriment la relation établie entre l'adjectif défini et un nom (duquel il provient) ou un verbe (exprimant souvent l'action ou état réalisé/subi par le nom qualifié par l'adjectif).

La reconnaissance de certaines structures porteuses de rapports synonymiques ou antonymiques entre concepts se révèle aussi puissante pour l'établissement de ces liens dans la représentation sémantique des définitions.

Comme on a dit précédemment, un analyseur syntaxique-sémantique robuste va être un outil indispensable pour ce passage des définitions aux structures sémantiques. Nous sommes actuellement en train de construire un *«parser»* fondé sur une hierarchie de schémas (Alshawi, 1987, 1989) qui va nous permettre d'obtenir une représentation au moins partielle des textes de définition analysés.

3.6. Le formalisme de représentation

Le dictionnaire va être représenté comme un réseau de concepts reliés à la manière de la *mémoire sémantique* de M. R. Quillian (1968). Les rapports entre concepts qui vont être représentés seront ceux dont on a parlé dans la section precédente.

Quillian établit deux types de relations entre les éléments constitutifs d'une définition:

— *intraplane liks* ou relations horizontales
— *interplane links* ou relations verticales

Les relations que nous appellons horizontales sont des relations établies au niveau où se trouvent les constituants de la définition, relations entre le *type* ou concept défini et les éléments du texte de la définition; de même on inclut ici les liens entre les éléments constitutifs du texte de la définition. On va prendre en compte ici les relations taxonomiques (hyperonymie, hyponymie), synonymiques, antonymiques et aussi les relations casuelles (agent, objet, etc.).

Les relations verticales sont des relations établies entre les differents *token* ou occurrences (instances) apparaîssant dans les définitions et les *types* ou noeuds de la mémoire sémantique qui seront dans notre cas les sens ou concepts représentés dans le dictionnaire. Ces *token* représentent des instances particulières des *types* correspondants. Pour cette représentation qui est implémentée comme un réseau de structures du style frame, nous utilisons deux types de constructions de base:

CONCEPT-TYPE: représente un concept ou sens du dictionnaire (le dictionnaire peut être vu comme un réseau de *concept-type* qui est accessible précisement au travers de ces concepts).
CONFIGURATION: représente une instance ou occurrence particulière d'un concept dans un texte concret de définition.

Basiquement, la définition d'un concept est représentée par:

(a) la configuration ou l'instance du concept hyperonyme du défini (définition classique du type *genus-differentia*)
ou (b) au moyen de concepts liés au défini par des relations lexicales de type synonymique, antonymique, etc.
ou (c) au moyen d'autres relations obtenues à partir de l'étude réalisée sur le métalangage définitoire des dictionnaires (mots-fonction): *ACTION-DE, PARTIE-DE, CE-QUI,* etc.

L'attribut le plus important dans le frame qui représente un concept-type A est celui qui contient une *référence* à un autre concept B ou à une configuration instance d'un autre concept B. C'est ce dernier cas le cas des définitions classiques où une phrase définit le concept, ayant cette phrase un noyau (B) qui est l'hyperonyme ou genus du concept A. Par contre, dans les cas où les concepts A et B sont réliés directement la définition utilise des relations lexicales particulières: synonymie, antonymie, ou d'autres relations spécifiques propres au métalangage dictionnarial.

Comme on l'a dit ci-dessus, les configurations représentent des occurrences particulières des concepts-type dans des phrases de définition concrètes. Elles ont un slot caractéristique nommé *modalité* qui exprime des circonstances concrètes de la configuration dans la définition en question: détermination, quantification, mode verbal, aspect, temps, personne, etc. Mais bien sûr, l'attribut le plus important de toute configuration est le concept-type dont elle est instance.

On trouve 4 types de configurations dépendant de la catégorie du noyau correspondant. Les configurations nominales sont utilisées pour représenter des structures du type groupe nominal et sont des instances de concepts nominaux ou pronominaux. Parmi les slots propres aux configurations nominales on trouve: CARACTÉRISTIQUE, PRÉDICATION, ORIGINE, BUT, POSSESSEUR, etc. mais on peut trouver aussi des mots-fonction comme ACTION-DE, CE-QUI, etc. Les configurations verbales représentant des instances de concepts de catégorie verbale possédent des slots «casuels»: AGENT, MANIÈRE, INSTRUMENTAL, etc. Les configurations adjectivales possèdent très souvent un slot GRADUATEUR qui peut être rempli avec un concept adverbial ou même adjectival et des slots dérivés du métalangage comme ÉTAT-DE, QUI, RELATIF-À, etc. De même, les configurations adverbiales représentant des instances de concepts d'adverbe ont éventuellement aussi un slot GRADUATEUR.

L'utilisation d'un logiciel de représentation à base de frames (KEE-Intellicorp) avec des caractéristiques orientées objet, gérant les hiérarchies entre objets, nous permet d'établir des hiérarchies entre concepts, et donc, d'exploiter des propriétés héritées des niveaux supérieurs aux niveaux inférieurs. Ce logiciel nous permet aussi de spécifier le domain de chacun de ces slots, le type des valeurs qui vont les remplir, en assurant ainsi la consistence des données introduites.

La méthode d'accès au dictionnaire tient compte des concepts à plusieurs acceptions permettant à l'utilisateur d'accéder à un sens particulier ou à tous les sens d'un mot. De façon interne le système ne considère qu'un concept par sens.

L'aspect déductif ou inférentiel du système va se fonder sur des règles du type de celles des systèmes experts. Pour le moment nous n'avons commencé à caractériser que certains aspects sous la forme de règles déductives aptes à dériver des faits à partir des faits explicites contenus dans le dictionnaire (de plus, un système de main-

tien de la vérité, TMS, du logiciel utilisé s'occupe de la gestion de ces faits pendant le fonctionnement du système). Les aspects couverts actuellement sont les suivants:

— transitivité des relations taxonomiques: hyponymie et hyperonymie
— synonymie et antonymie: symetrie et transitivité
— synonymes des hyperonymes
— hyperonymes des synonymes

4. Conclusion et perspectives

Dans ce travail nous avons décrit sommairement ce que nous appelons le *dictionnaire intelligent d'aide à la compréhension.* La première phase du projet a consisté à étudier un dictionnaire réel dans le but d'examiner de quelle sorte est l'information qu'il contient. On s'est fixé surtout sur la composante sémantique du dictionnaire et on a procédé à l'étude de la syntaxe des textes de définition en dégageant certaines structures et mots —*mots-fonction*—. On va utiliser ces structures comme base des constructions sémantiques qui vont composer le dictionnaire.

Pour la représentation sémantique à base de frames on profite des possibilités offertes par les environments d'intelligence artificielle —héritage, programmation orientée object, etc.—, dans le but de l'implantation informatique d'un prototype de cette interface intelligente.

Ensuite on va entreprendre des études sur la fonctionnalité d'utilisation des dictionnaires qui vont nous offrir une idée des caractéristiques du processus d'utilisation realisé lors de la consultation d'un dictionnaire. Cette caractérisation va spécifier l'aspect inférentiel du système. On prevoit de travailler aussi sur cette idée dans le but de découvrir, caractériser et classer les fonctions d'utilisation dans un environnement d'aide à la traduction où l'usager ne se sert pas seulement d'un dictionnaire monolingue mais d'un vaste ensemble de dictionnaires plurilingues, monolingues, techniques, généraux, synonymiques, etc.

5. Références

ALSHAWI, H., «Processing dictionary definitions with Phrasal Pattern Hierarchies», *Computational Linguistcs* 13, 3-4, 195-202, 1987.

ALSHAWI, H., «Analysing dictionary definitions» in B. Boguraev, T. Briscoe eds., 153-169, *Computational Lexicography for Natural Language Processing.* New York: Longman, 1989.

AMSLER, R.A., *The Structure of the Merriam-Webster Pocket Dictionary*, Ph.D. Dissertation, Computer Science. University of Texas, Austin, 1980.

AMSLER, R.A. «A Taxonomy for English Nouns and Verbs», *Proc. 19th Annual Meeting ACL*, 133-138, 1981.

AMSLER, R.A. «Lexical Knowledge Bases», *Proc. COLING* (Standford Univ.), 458-459, 1984.

ARANGO GAVIRÍA, G., *Une approche pour amorcer le processus de compréhension et d'utilisation du sens des mots en langage naturel.* Thèse de 3e cycle (Paris VI). Publications du Groupe de Recherche Claude François Picard, 1983.

BINOT, J.L., JENSEN, K., «A semantic expert using an on-line standard dictionary», *Proc. IJCAI*, 709-714, 1987.

BOGURAEV, B., BRISCOE, T., eds. *Computational Lexicography for Natural Language Processing.* New York: Longman, 1989.

BRUSTKERN, J., HESS, K.D. «The BONNLEX lexicon system» in J. Goetschalckx, L. Rolling eds., 33-40, *Lexicography in the electronic age*. Luxembourg: North-Holland, 1982.

BYRD, R.J., CALZOLARI, N., CHODOROW, M.S., KLAVANS, J.L., NEFF, M.S., RIZK, O.A. «Tools and Methods for Computational Lexicography» *Computational Linguistics* 13, 3-4, 219-240, 1987.

CALZOLARI, N., PECCHIA, L., ZAMPOLLI, A. «Working on the italian machine dictionary: a semantic approach» in A. ZAMPOLLINI, N. CALZOLARI eds. *Proceedings of the International Conference on Computational Linguistics* (Pisa, 1973). Vol. 37, 49-69, 1980.

CALZOLARI, N. «Machine-readable dictionaries, lexical data bases and the lexical system», *Proc. COLING* (Standford Univ.), p. 460, 1984a.

CALZOLARI, N. «Detecting patterns in a lexical data base», *Proc. COLING* (Standford Univ.), 170-173, 1984b.

CHODOROW, M.S., BYRD, R.J. «Extracting semantic hierarchies from a large on-line dictionary», *Proc. ACL*, 299-304, 1985.

CHODOROW, M.S., RAVIN, Y., SACHAR, H.E. «A tool for investigating the synonymy relation in a sense disambiguated thesaurus», *Proc. 2nd conference on Applied Natural Language Processing* (Austin), 144-153, 1988.

CHOURAQUI, E., GODBERT, E. «Représentation des descriptions définies dans un réseau sémantique», *Actes 7ème Congrès Reconnaissance des Formes et Intelligence Artificielle* (AFCET-INRIA, Paris), 855-868, 1989.

FERRARI, G. «Dictionnaire automatique et dictionnaire-machine: une hypothèse» in A. Zampolli, N. Calzolari eds. *Proceedings of the International Conference on Computational Linguistics* (Pisa, 1973). Vol. 36, 257-262, 1980.

VAN DEN HURK, I., MEIJS, W. «The dictionary as a corpus: analyzing LDOCE's definition-language», *Corpus Linguistics II*, 99-125.

LITKOWSKY, K.C. «Models of the semantic structure of dictionaries», *American Journal of Computational Linguistics*, Mf. 81, 25-74, 1978.

MARKOWITZ, J., AHLSWEDE, T., EVENS, M. «Semantically significant patterns in dictionary definitions», *Proc. 24th Annual Meeting ACL* (New York), 112-119, 1986.

MICHIELS, A., NOËL, J. «Approaches to thesaurus production» in J. Horecky ed., 227-232, *COLING 82*. Academia: North-Holland, 1982a.

MICHIELS, A., MULLENDERS, J., NOEL, T. «The LONGMAN-LIEGE project» in J. Goetschalckx, L. Rolling eds., 201-210, *Lexicography in the electronic age*. Luxembourg: North-Holland, 1982b.

NAGAO, M., TSUJII, J., UEDA, Y. et TAKIYAMA, M., «An attempt to computerize dictionary data bases» in J. Goetschalckx, L. Rolling eds., 51-73, *Lexicography in the electronic age*. Luxembourg: North-Holland, 1982.

PAPEGAAIJ, B.C., SADLER, V., WITKAM, A.P.M., «Experiments with an MT-directed lexical knowledge bank», *Proc. COLING* (Bonn), 432-434, 1986.

PAZIENZA, M.T., VELARDI, P. «A structured representation of word-senses for semantic analysis», *Proc. 3rd European Conference ACL* (Copenhaguen), 249-257, 1987.

QUILLIAN, M.R., «Semantic Memory» in M. Minsky ed., 227-270, *Semantic Information Processing*. Cambridge (Mass.): MIT Press, 1968.

TSURUMARU, H., HITAKA, T., YOSHIDA, S., «An attempt to automatic thesaurus construction from an ordinary japanese language dictionary», *Proc. COLING* (Bonmn), 445-447, 1986.

VOSSEN, P., MEIJS, W., DEN BROEDER, M. «Meaning and structure in dictionary definitions» in B. Boguraev, T. Briscoe eds., 171-192, *Computational Lexicography for Natural Language Processing*. New York: Longman, 1989.

WINSTON, M., CHAFFIN, R., HERMANN, D. «A taxonomy of part-role relation», *Cognitive Science*, vol. 11, 417-444, 1987.

YOSHIDA, S., TSURUMARU, H., HITAKA, T., «Man-assisted machine construction of a semantic dictionary for natural language processing» in J. Horecky ed., 419-424, *COLING 82*. Academia: North-Holland, 1982.